УДК 004.912

О.В. Лозинська, М.В. Давидов, В.В. Пасічник
Національний університет «Львівська політехніка», Україна
вул. С. Бандери, 12, м. Львів, 79000

ТРАНСФОРМАЦІЯ ДЕРЕВ ГРАМАТИКИ СКЛАДОВИХ У ДЕРЕВА ГРАМАТИКИ ЗАЛЕЖНОСТЕЙ ДЛЯ ГРАМАТИЧНОГО РОЗБОРУ УКРАЇНСЬКИХ РЕЧЕНЬ

O.V. Lozynska, M.V. Davydov, V.V. Pasichnyk
Lviv Polytechnic National University, Ukraine
S. Bandery Str., 12, Lviv, 79000

TRANSFORMATION OF CONSTITUENCY TREES TO DEPENDENCY TREES FOR PARSING UKRAINIAN SENTENCES

У статті розглянуто метод трансформації дерев граматики складових у дерева граматики залежностей, який використовується для перекладу речень української словесної мови у речення анотованої української жестової мови. Зроблено граматичний розбір корпусу речень «Українська словесна мова» та побудовано дерева граматичного розбору цих речень (дерева граматики складових). Описано кроки алгоритму трансформації дерев граматики складових у дерева граматики залежностей для речень української словесної мови.
Ключові слова: граматичний розбір, дерево синтаксичного розбору, трансформація дерев, граматика складових, граматика залежностей, машинний переклад.
The method of converting constituency structure to dependency structures that using for translation of Ukrainian spoken language into annotated Ukrainian sign language is considered in the article. The parsing of sentences of corpora "Ukrainian Spoken Language" are made and the parsing trees of this sentences are built. The steps of the transformation algorithm of Ukrainian spoken language are described.
Keywords: parsing, parsing tree, trees transformation, dependency grammar, constituency grammar, machine translation.

Вступ

Розроблення та вивчення способів комп’ютерного опрацювання речень та побудова їх синтаксичної структури є актуальним завданням сьогодення. Для опису синтаксичної структури речення можна або виділити в ньому складові – групи слів, що функціонують як цілісні синтаксичні одиниці, або вказати для кожного слова ті слова, які йому безпосередньо підпорядковані. У першому випадку використовується граматика складових і будується дерево складових, у другому випадку використовується граматика залежностей і, відповідно, будується дерево залежностей.

Побудова та комп’ютерне представлення синтаксичної структури речень часто використовується у системах машинного перекладу на основі правил та на основі онтологій. На вхід системи машинного перекладу подається речення, яке проходить граматичний аналіз з використанням граматики складових. В результаті синтаксичного аналізу будується дерево синтаксичного розбору цього речення (дерево граматики складових). Використання граматики складових вимагає перетворення дерева розбору у граматику залежностей для подальшого застосування системи правил перекладу.

Постановка проблеми

Машинний переклад української словесної мови (УСМ) на анотовану українську жестову (УЖМ) та навпаки поділяється на декілька етапів, а саме: граматичний розбір речень (в результаті чого будується дерево складових), трансформація дерева складових вхідного речення у дерево залежностей згідно з алгоритмом трансформації, перетворення дерева залежностей у речення анотованої УЖМ з використанням правил перекладу на жестову мову, правил порядку слів у реченнях відповідно до граматики УЖМ.

Оскільки під час граматичного аналізу речень УСМ будується дерево граматики складових, потрібно розробити алгоритм трансформації дерев граматики складових (граматичного розбору) у дерева граматики залежностей.

Для практичного втілення цього алгоритму необхідно описати правила визначення головної складової у дереві граматики складових та описати послідовність кроків для трансформації дерева складових у дерева залежностей для усіх типів речень.

Авторами досліджено застосовність розробленого алгоритму трансформації дерев граматики складових у дерева граматики залежностей на корпусі речень «Українська словесна мова».

Аналіз останніх досліджень та публікацій

Автоматичний аналіз речень все частіше застосовується для розв‘язання широкого кола лінгвістичних задач, таких як машинний переклад, видобування і пошук інформації [1]. Існує два основних підходи до аналізу структури речення: підхід на основі граматики залежностей і підхід на основі граматики складових.

Результатом синтаксичного аналізу речення є дерево синтаксичного розбору, яке відповідає граматиці залежностей. Дерева граматики залежностей переважно використовуються для мов із вільним порядком слів (наприклад, в українській), а дерева граматики складових – для мов з строго визначеним порядком слів (наприклад, в англійській, українській жестовій мові).

Перетворення дерева граматичного розбору із застосуванням граматики складових до дерева синтаксичного розбору вимагає застосування додаткових алгоритмів, які визначають головні складові мовних конструкцій та будують на їх основі дерева розбору.

Синтаксичний аналіз речень української словесної мови у вигляді дерев залежностей досліджено у роботах Н. Дарчук [2], М. Лангенбах [3] та ін. У роботі [2] розроблено програмне забезпечення, яке протестовано на корпусі української мови, який містить 650 тис. речень. У роботі [3] наведено алгоритм автоматичного моделювання структури речення в термінах граматики залежностей та описано формалізацію правил установлення зв’язків у реченні та їх автоматизацію. Автором наведено основні переваги та недоліки обраного алгоритму. Проте у відкритому доступі відсутні тестові корпуси та програмна реалізація наведених алгоритмів.

У роботі [4] російських вчених А. Антонова та ін. розроблено синтаксичний аналізатор для російської та англійської мов, використовуючи граматику залежностей. На вхід синтаксичного аналізатора подається файл з текстом російською або англійською мовою. Опрацювання тексту складається з таких етапів: розбиття тексту на речення і слова, морфологічний розбір, синтаксичний розбір, інтерпретація результатів синтаксичного розбору.

У роботі [5] наведено три алгоритми трансформації дерева залежностей у дерево складових для англійської мови та проведено їх застосовність на корпусі речень Penn Treebank [6], для яких побудовано синтаксичну структуру розбору. Найкращий результат трансформації дерева отримано за допомогою алгоритму №3.

Алгоритм трансформації дерева складових у дерево залежностей для англійської мови описано у роботі [7], який дав змогу зменшити кількість помилок розбору на 23%.

У роботі [8] подано новий алгоритм для перетворення дерева залежностей у дерево складових. Даний алгоритм досягає 90,4% розбору для речень англійською мовою і 82,4% розбору для речень китайською мовою.

Основні правила визначення головної складової у дереві складових описані у роботі [1]. Наведено алгоритм трансформації дерева складових у дерево залежностей та

проведено тестування даного алгоритму на 232 реченнях. Алгоритм трансформації позначає відношення між різними складовими дерева складових і перетворює його у дерево залежностей. Запропоновані іноземними вченими алгоритми не можуть бути застосовані для речень української мови, оскільки не враховують специфіку граматичного розбору флективних мов, до яких відноситься українська мова.

Алгоритм трансформації дерева складових у дерево залежностей

Після граматичного розбору речення виконують трансформацію дерева складових цього речення у дерево залежностей згідно з алгоритмом трансформації, поданого у [1]. Як для дерев граматики залежностей, так і для дерев граматики складових важливим є поняття «головна складова» (англ. «head»). Для дерев граматики складових головною складовою позначається головне слово у виразі і від цього слова залежать усі інші слова виразу. Ядром алгоритму є визначення головної складової кожного виразу для дерев граматики складових і встановлення зв'язку з головною складовою його батьківського вузла. Головною складовою для кожної іменникової групи (ГРУПА ІМЕННИКА) є вузол ІМЕННИК або ЗАЙМЕННИК в цьому вузлі ГРУПА ІМЕННИКА, а головною складовою для дієслівної групи (ГРУПА ПРИСУДКА) є ДІЄСЛОВО. Головною складовою для вузла ОСНОВА РЕЧЕННЯ є головна складова ГРУПА ПРИСУДКА, якщо ОСНОВА РЕЧЕННЯ є простим реченням, та головна складова ГРУПА ПРИСУДКА головного речення, якщо ОСНОВА РЕЧЕННЯ є складним реченням.

Крім того, для кожного вузла дерева складових речення визначається семантичний атрибут. Зазвичай цей атрибут копіюється з головної складової піддерева дерева складових речення. Винятком можуть бути усталені вирази, в яких значення слів не відповідають значенню цілого виразу.

Правила визначення головної складової в деревах граматики складових такі:

1) головна складова вузла ГРУПА ПРИСУДКА або ОСНОВА РЕЧЕННЯ є кореневим вузлом (коренем) у дереві граматики залежностей;

2) якщо вузол ОСНОВА РЕЧЕННЯ є батьком ГРУПА ПРИСУДКА, то всі ГРУПА ІМЕННИКА, які є нащадками вузла ОСНОВА РЕЧЕННЯ, також є залежними від цього кореня;

3) головними складовими вузлів ГРУПА ІМЕННИКА, ПРЯМИЙ ДОДАТОК, ДОДАТОК, є вузол ІМЕННИК або ЗАЙМЕННИК. Усі решта вузлів, які входять у ГРУПУ ІМЕННИКА, ПРЯМОГО ДОДАТКА, ДОДАТКА, зокрема ПРИКМЕТНИКИ є залежними від них;

4) головною складовою вузла ГРУПА ПРИСУДКА є вузол ДІЄСЛОВО. Якщо вузол ГРУПА ПРИСУДКА містить ДОПОМІЖНЕ ДІЄСЛОВО, то воно залежить від ДІЄСЛОВА;

5) головною складовою вузлів ОБСТАВИНА МІСЦЯ, ОБСТАВИНА ЧАСУ, ОБСТАВИНА МЕТИ, ОБСТАВИНА ПРИЧИНИ, ОБСТАВИНА УМОВИ, ОБСТАВИНА СПОСОБУ ДІЇ, ОБСТАВИНА ДОПУСКУ, ОБСТАВИНА МІРИ, які виражені:

а) іменником з прийменником є ПРИЙМЕННИК;

б) дієприслівниковим зворотом є ДІЄПРИСЛІВНИК.

Блок-схему алгоритму трансформації дерева складових у дерево залежностей зображено на рис. 1.



Рис. 1. Блок-схема алгоритму трансформації дерева складових у дерево залежностей

Алгоритм трансформації визначає відношення нащадок-батько для вузлів дерева граматики складових і перетворює його на дерево граматики залежностей. Просуваючись по відношенням нащадок-батько від вузлів-листків до вузла-кореня дерева, ми можемо позначити кожен вузол тегом «головна складова». Цей алгоритм дає змогу визначити головну складову для кожного вузла дерева і відповідно залежну складову цього вузла, яка підпорядковується головній складовій. Просуваючись по дереву складових вверх, ми отримаємо головну складову всього речення (кореня дерева).

Для кожної неузгодженої залежності здійснюється пошук вшир у дереві складових, починаючи від її батьківського вузла, для того щоб знайти слово, від якого неузгоджена складова може бути залежна.

Наведений алгоритм дає змогу перетворити дерева граматичного розбору, які отримані за допомогою парсеру UkrParser [9], на дерева залежностей. Після цього

можна перекладати речення української словесної мови на українську жестову мову та навпаки, використовуючи основні правила граматики [10].

На рис. 2 подано дерево складових, яке будується в процесі роботи парсеру UkrParser, а на рис. 3 зображено результат роботи алгоритму та його основні етапи (визначення головної складової кореня дерева) для речення: «Гарні студенти прийшли сьогодні на пари». Визначення головної складової речення зображено потовщеними лініями.

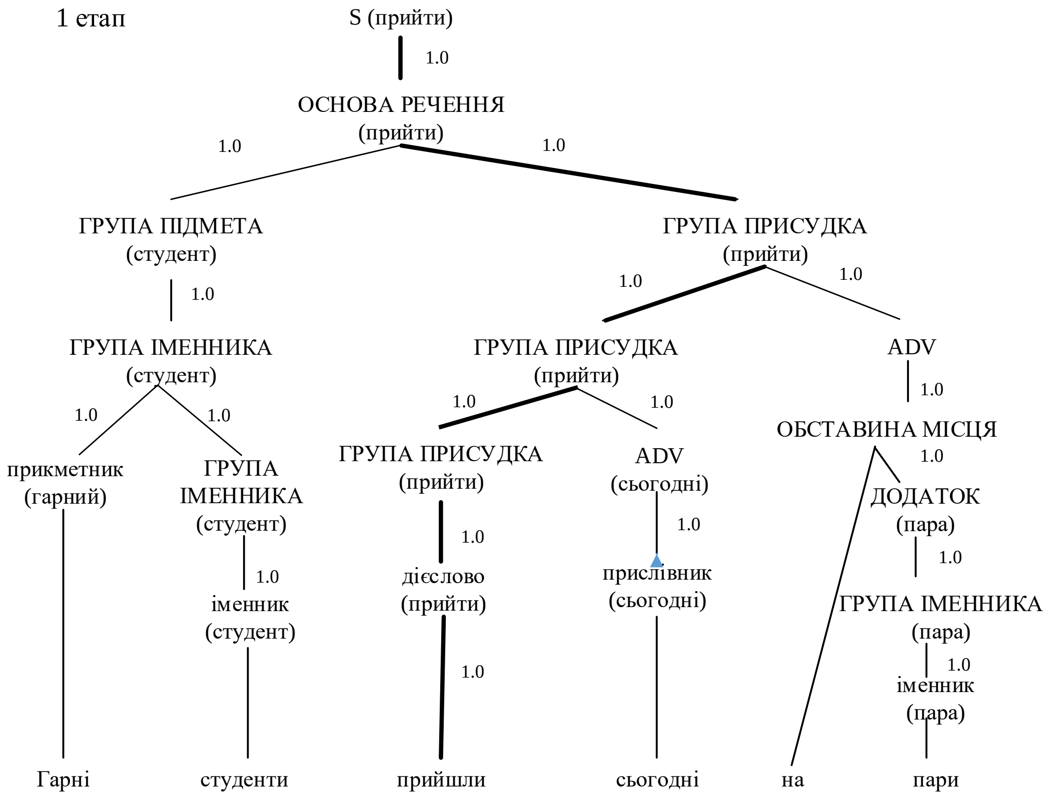


Рис. 2. Граматичний розбір та дерево складових речення
«Гарні студенти прийшли сьогодні на пари»

На перщому етапі визначається корінь дерева залежностей, який є головною складовою ГРУПИ ПРИСУДКА і є семантичним атрибутом – це дієслово ПРИЙШЛИ (ПРИЙТИ). На другому етапі шукаються залежні від дієслова вузли – в даному прикладі це прислівник СЬОГОДНІ. І це слово-вузол ставиться на рівень нижче від кореня дерева залежностей. На цьому ж етапі шукаються ГРУПА ПІДМЕТА, всі ДОДАТКИ (прямі чи непрямі) та ОБСТАВИНИ. У них визначаються головні складові і записуються на рівень нижче від кореня дерева залежностей. Наприклад, для речення, що зображене на рис. 3:

1) головна складова вузла ГРУПИ ПІДМЕТА – іменник СТУДЕНТИ;

2) головна складова вузла ОБСТАВИНА МІСЦЯ – прийменник НА.

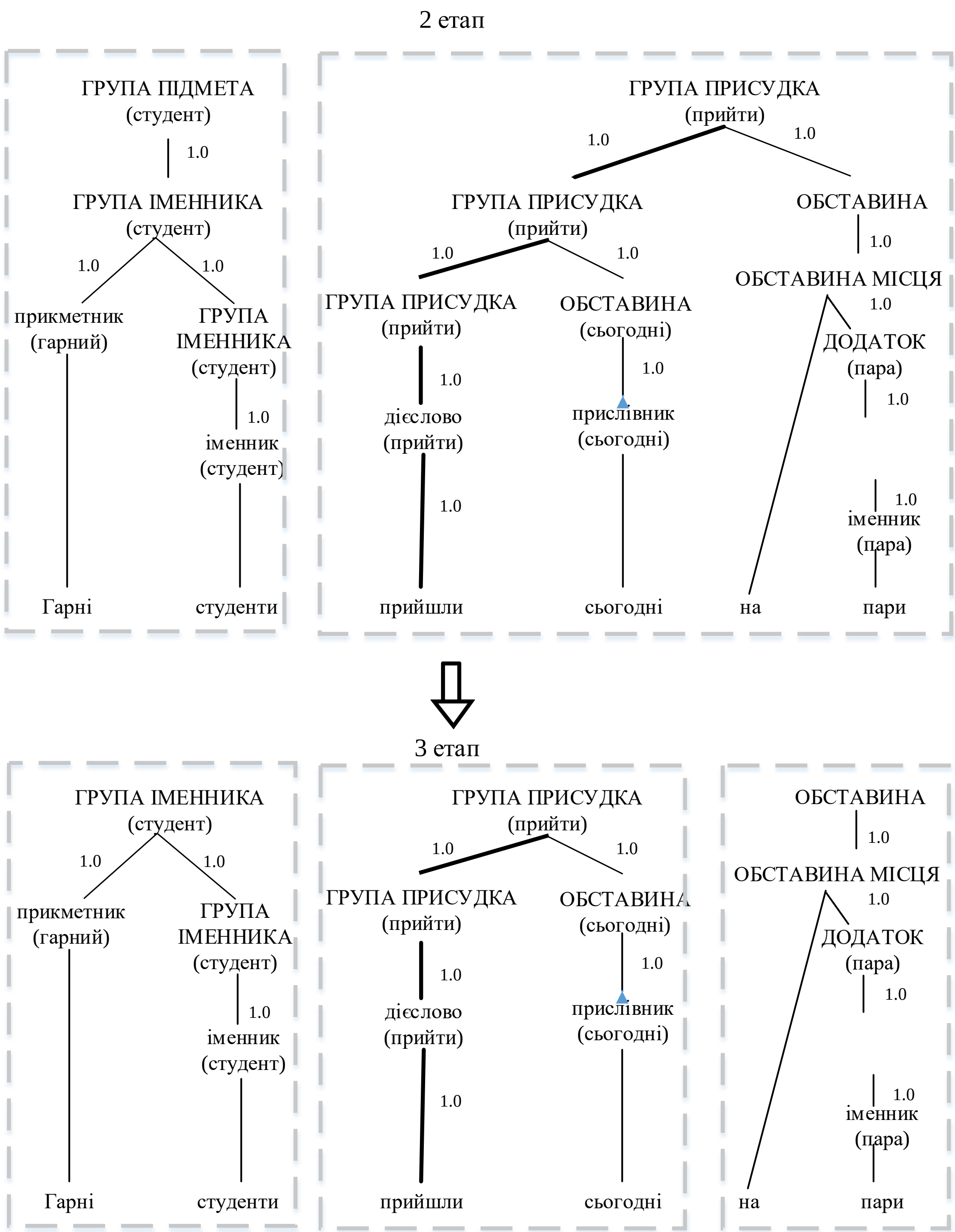


Рис. 3. Етапи визначення головної складової у дереві складових для речення «Гарні студенти прийшли сьогодні на пари»

На третьому етапі шукаються усі залежні слова у ГРУПІ ПІДМЕТА, ДОДАТКА та ОБСТАВИНИ і ставляться на рівень нижче від їхніх вузлів-батьків. В даному прикладі залежний вузол у ГРУПІ ПІДМЕТА – це прикметник ГАРНІ, а залежний вузол у ОБСТАВИНА МІСЦЯ – це іменник ПАРИ.

Так, можна побудувати дерева граматики залежностей. Згідно з описаним алгоритмом дерево граматики залежностей для речення «Гарні студенти прийшли сьогодні на пари» зображено на рис. 4.

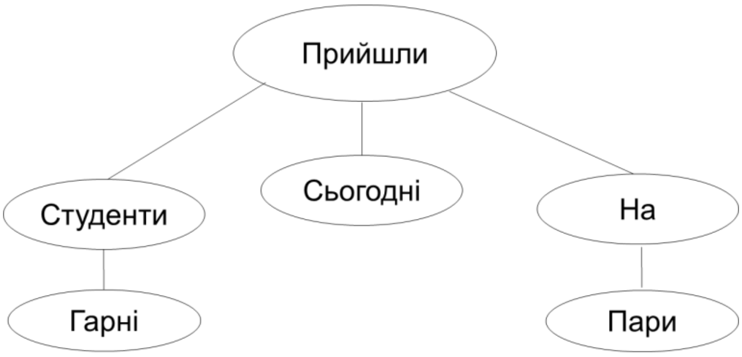


Рис. 4. Дерево граматики залежностей для речення
«Гарні студенти прийшли сьогодні на пари»

Оскільки порядок слів в українській словесній мові не строго визначений, то речення «Гарні студенти прийшли сьогодні на пари», «Студенти гарні сьогодні на пари прийшли» та «Студенти сьогодні гарні на пари прийшли» будуть однакові за змістом.

Розглянено приклад речення з незвичайним порядком слів «Моя донька у садок ходить дитячий», дерево складових якого зображено на рис. 5.

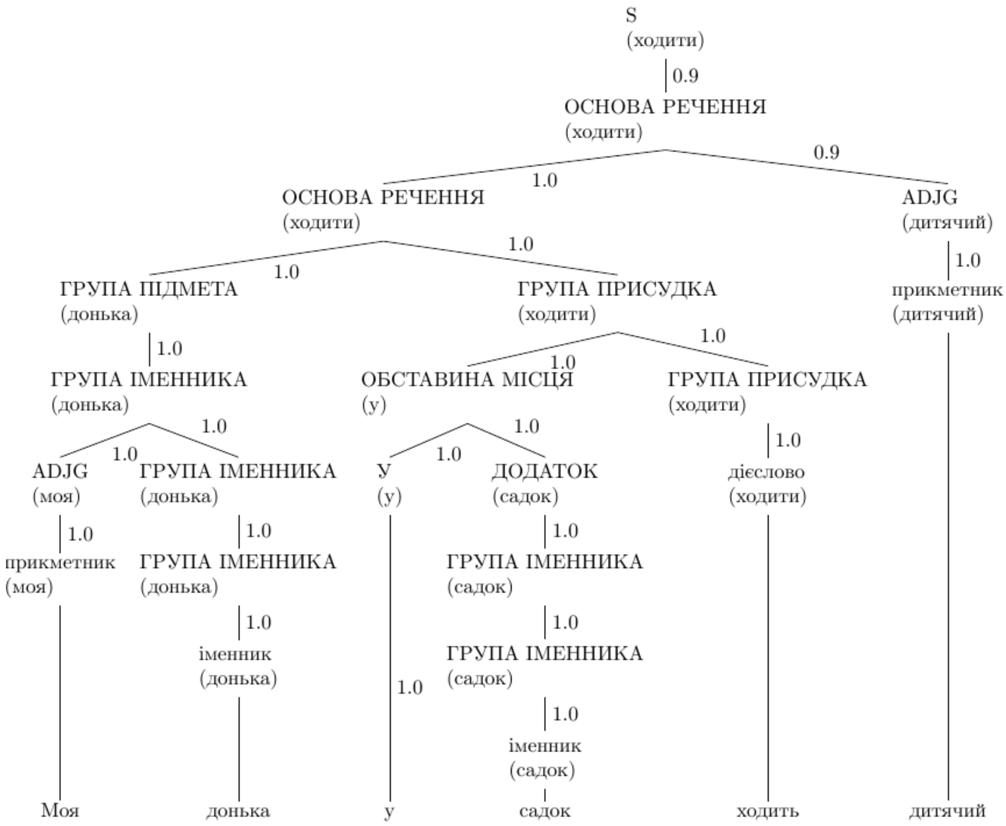


Рис. 5. Дерево складових речення з незвичайним порядком слів
«Моя донька у садок ходить дитячий»

У цьому реченні без використання наведеного алгоритму визначення головної складової і трансформації у дерево залежностей слово «дитячий» є залежним від слова «ходить» (рис. 6).

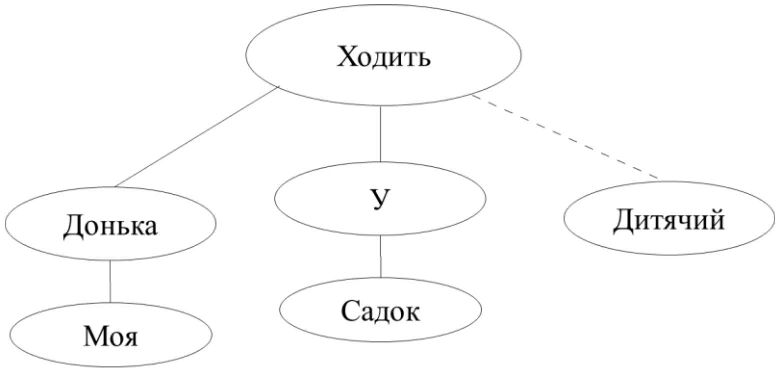


Рис. 6. Некоректне дерево граматики залежностей для речення
«Моя донька у садок ходить дитячий»

Враховуючи правила української мови про іменникове словосполучення, прикметник «дитячий» є залежним від іменника «садок». Оскільки у алгоритмі трансформації на останньому кроці передбачено виправлення неузгоджених зв’язків методом пошуку вшир у дереві залежностей, то для слова «дитячий» буде знайдено найближче слово «садок», яке може бути батьківським до слова «дитячий» у дереві залежностей. Коректне дерево граматики залежностей для речення «Моя донька у садок ходить дитячий» зображене на рис. 7.

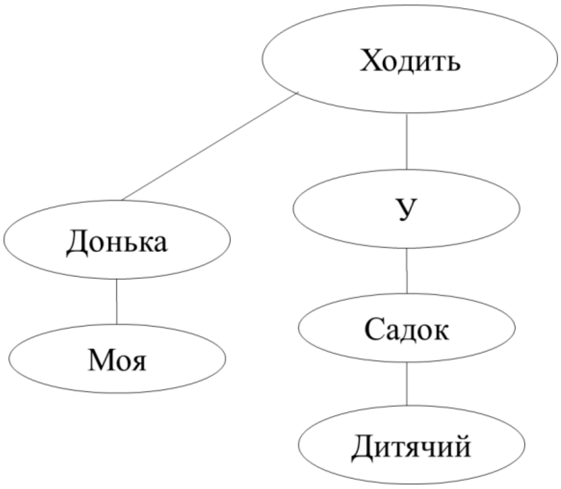


Рис. 7. Коректне дерево граматики залежностей для речення
«Моя донька у садок ходить дитячий»

Аналіз результатів

Розроблений алгоритм перетворення дерева складових на дерево залежностей протестований на 300 реченнях корпусу «Українська словесна мова». Результати тестування алгоритму трансформації дерева складових у дерево залежностей показали високу ефективність застосування цього алгоритму. Правильно перетворено 92% речень корпусу, що на 3% більше, ніж при перетворенні без використання останнього

кроку алгоритму виправлення неузгоджених зв'язків методом пошуку вшир у дереві залежностей.

Алгоритм визначення головної складової у дереві складових добре працює на тих реченнях УСМ, у яких чітко визначено порядок слів. Проте, є випадки, що під час розбору не завжди можна прикріпити залежні слова до головної складової. Наприклад, коли у реченні присудок є складеним.

Висновки

У статті досліджено метод трансформації дерев граматики складових у дерева граматики залежностей, який використовується для перекладу речень української словесної мови у речення анотованої української жестової мови. Описано кроки алгоритму трансформації для речень української словесної мови.

Проведені дослідження застосування алгоритму трансформації на реченнях української словесної мови показали високу ефективність цього алгоритму (92 % правильно перетворених речень) та можливість його використання в системах машинного перекладу. Також зазначено недоліки алгоритму, а саме некоректне перетворення речень, у яких не має чітко визначеного порядку слів. Наступним дослідженням стане вдосконалення алгоритму трансформації для підвищення його ефективності.

Література

1. Goyal P. Converting Phrase Structures to Dependency Structures in Sanskrit / P. Goyal, A. Kulkarni // Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, 2014. – P. 1834–1843.
2. Дарчук Н.П. Автоматичний синтаксичний аналіз текстів корпусу української мови / Н.П. Дарчук // Українське мовознавство. – Київ. нац. ун-т ім. Т. Шевченка, 2013. – № 43. – С. 11–19.
3. Лангенбах М.О. Автоматичний синтаксичний аналіз речення за принципами граматики залежностей / М.О. Лангенбах // Науковий вісник Східноєвропейського національного університету імені Лесі Українки, 2015. – Т. 3. – С. 249-254.
4. Антонова А.А. Синтаксический анализатор для русского и английского языков / А.А. Антонова, А.В. Мисюрев // Сб. трудов ИСА РАН / Под ред. В.Л. Арлазарова и Н.Е. Емельянова. – М.: УРСС, 2007. – С .329 – 337.
5. Xia F. Converting dependency structures to phrase structures / F. Xia, M. Palmer // In Proceedings of the first international conference on Human language technology research, Association for Computational Linguistics, 2001. – P. 1–5.
6. Marcus M. Building a Large Annotated Corpus of English: the Penn Treebank Computational Lingustics / M. Marcus, B. Santorini, M.A. Marcinkiewicz // Computational Linguistics, Vol. 19, Issue 2, 1993.– P. 313-330.
7. Johansson R. Extended Constituent-to-dependency Conversion for English / R. Johansson, P. Nugues // In Proceedings of NODALIDA, Tartu, Estonia, 2007. – P. 105–112.
8. Kong L. Transforming Dependencies into Phrase Structures / L. Kong, A.M. Rush, N.A. Smith // In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT 2015), Denver, CO, 2015. – P. 788-798.
9. Проект «UkrParser» [Електр. ресурс]. – Режим доступу: https://github.com/mdavydov/UkrParser
10. Лозинська О.В. Машинний переклад на основі правил для перекладу на українську жестову мову / О.В. Лозинська, М.В. Давидов, В.В. Пасічник // Міжнародний науково-технічний журнал "Інформаційні технології та комп'ютерна інженерія". – Вінниця, 2014. – С. 11–17.

Literatura

1. Goyal P. Converting Phrase Structures to Dependency Structures in Sanskrit / P. Goyal, A. Kulkarni // Proceedings of COLING 2014, Technical Papers, Dublin, Ireland, 2014. – P. 1834–1843.
2. Darchuk N.P. Avtomatychnyi syntaksychnyi analiz tekstiv korpusu ukrainskoi movy / N.P. Darchuk // Ukrainske movoznavstvo. – Kyiv. nats. un-t im. T. Shevchenka, 2013. – № 43. – S. 11–19.

3. Lanhenbakh M.O. Avtomatychnyi syntaksychnyi analiz rechennia za pryntsypamy hramatyky zalezhnostei / M.O. Lanhenbakh // Naukovyi visnyk Skhidnoievropeiskoho natsionalnoho universytetu imeni Lesi Ukrainky, 2015. – T. 3. – S. 249-254.
4. Antonova A.A. Syntaksycheskyi analyzator dlia russkoho y anhlyiskoho yazыkov / A.A. Antonova, A.V. Mysiurev // Sb. trudov YSA RAN / Pod red. V.L. Arlazarova y N.E. Emelianova. – M.: URSS, 2007. – S. 329 – 337.
5. Xia F. Converting dependency structures to phrase structures / F. Xia, M. Palmer // In Proceedings of the first international conference on Human language technology research, Association for Computational Linguistics, 2001. – P. 1–5.
6. Marcus M. Building a Large Annotated Corpus of English: the Penn Treebank Computational Lingustics / M. Marcus, B. Santorini, M.A. Marcinkiewicz // Computational Linguistics, Vol. 19, Issue 2, 1993.– P. 313-330.
7. Johansson R. Extended Constituent-to-dependency Conversion for English / R. Johansson, P. Nugues // In Proceedings of NODALIDA, Tartu, Estonia, 2007. – P. 105–112.
8. Kong L. Transforming Dependencies into Phrase Structures / L. Kong, A.M. Rush, N.A. Smith // In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT 2015), Denver, CO, 2015. – P. 788-798.
9. Proekt «UkrParser» [Elektr. resurs]. – Rezhym dostupu: https://github.com/mdavydov/UkrParser
10. Lozynska O.V. Mashynnyi pereklad na osnovi pravyl dlia perekladu na ukrainsku zhestovu movu / O.V. Lozynska, M.V. Davydov, V.V. Pasichnyk // Mizhnarodnyi naukovo-tekhnichnyi zhurnal “Informatsiini tekhnolohii ta komp’iuterna inzheneriia”. – Vinnytsia, 2014. – C. 11–17.

RESUME

O.V. Lozynska, M.V. Davydov, V.V. Pasichnyk

Transformation of constituency trees to dependency trees for parsing ukrainian sentences

The method of converting constituency structure to dependency structures that using for translation of Ukrainian spoken language into annotated Ukrainian sign language is considered in the article. The parsing of sentences of corpora "Ukrainian Spoken Language" are made and the parsing trees of this sentences are built. The steps of the transformation algorithm of Ukrainian spoken language are described.

The construction and computer representation of the syntactic structure of sentences are often used in machine translation systems based on rules and based on ontologies. As input, the machine translation system received the parsing sentence using constituency grammar. As a result of parsing sentence the parsing tree is built (constituency structure). For further application of the rules of translation from Ukrainian spoken language into annotated Ukrainian sign language it is need to converting constituency structure to dependency structures.

The algorithm of transformation of constituency structure to dependency structures tested on 300 sentences of corpora "Ukrainian Spoken Language". Test results of transformation has showed high efficiency of this algorithm (92% correctly transformed sentences).

The algorithm of transformation of constituency structure to dependency structures works well on those sentences, which clearly defined the word order. This problem can be solved by adding new rules for parsing sentences in which the word order is not strictly defined. A further step could be to extend the rules for parsing sentences to improve the efficiency of transformation algorithm.

Надійшла до редакції 31.10.2016